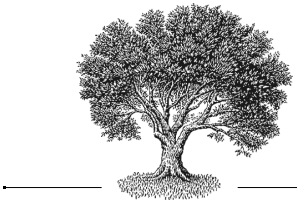

SlaVaComp: Konvertierungstool*

Программа конвертации шрифтов “SlaVaComp”

Simon Skilevic

Albert-Ludwigs-Universität Freiburg

Семён Сергеевич Шкилевич

Фрайбургский университет им. Альберта и Людвига

Zusammenfassung

Der vorliegende Beitrag informiert über ein Tool, das im Rahmen eines Freiburger Projekts zur historischen Korpuslinguistik entwickelt wurde und dazu dient, kirchenslavische Texte, die ohne Einsatz von Unicode digitalisiert wurden, ohne Verlust von Information bzw. Formatierung ins Unicode-Format zu überführen. Das Tool heißt SlaVaComp-Konvertierer. Es eignet sich für die Konvertierung aller idiosynkratischen Fonts und kann somit nicht nur in der Paläoslavistik, sondern in allen historisch arbeitenden Philologien eingesetzt werden.

Schlüsselwörter

Unicode, Konvertierungstool, kirchenslavische Alphabete und Schriften, Editionen kirchenslavischer Texte, Textwiederverwendung, Digitalisierung, idiosynkratische Fonts

* Der Beitrag ist im Rahmen des vom BMBF geförderten Projektes “SlaVaComp – COMPutergestützte Untersuchung von VAriabilität im KirchenSLAvischen” (FKZ: 01UG1251, Laufzeit: 15.01.2013–15.01.2016) unter der Leitung von Prof. Dr. J. Besters-Dilger und Prof. Dr. G. Schneider entstanden. Ich möchte mich an dieser Stelle herzlich bei allen ProjektmitarbeiterInnen für ihre Unterstützung bedanken. Mein besonderer Dank gilt Herrn Dr. Roman Krivko, Herrn Pino Marco Pizzo, M. A., Frau Dr. Irina Podtergera, Herrn Dr. Achim Rabus und Frau Viktoria Halapats, M. A., für die aufwändige Arbeit beim Ausfüllen der Konvertierungstabelle, für die Einblicke in die Wissenschaft von den Sprachen und für die Hilfe beim Finden und Beheben der unzähligen Fehler, die im Laufe des Entwicklungsprozesses auftraten.

Резюме

В статье сообщается о программе конвертации шрифтов разного формата в формат Юникод, разработанной в рамках фрайбургского проекта по исторической корпусной лингвистике. Программа обеспечивает конвертацию любых шрифтов в формат Юникод без потери информации, содержащейся в исконном файле, то есть с сохранением авторского форматирования. Название программы "SlaVaComp-Konvertierer / SlaVaComp-конвертор". Поскольку программа предназначена для конвертации любых, а не только церковнославянских, шрифтов, она может использоваться как палеославистами, так и представителями других историко-филологических дисциплин.

Ключевые слова

Юникод, программа конвертации, издания церковнославянских текстов, церковнославянские алфавиты и шрифты, повторное использование текстов, дигитализация, идиосинкратические шрифты

1. Einführung und Problembeschreibung

Editionen kirchenslavischer Texte werden in der Regel mit dem Einsatz von Non-Unicode-Fonts vorbereitet. Dies erschwert den Datenaustausch und die Wiederverwendung dieser Texte im digitalen Format erheblich: Sie sind nicht einheitlich eingegeben, sie werden auf dem Bildschirm falsch dargestellt oder lassen sich nicht lesen, zudem kommt es beim Übertragen in ein anderes Betriebssystem bzw. auf einen anderen Rechner häufig zu Kompatibilitätsproblemen, die die Funktionsfähigkeit von Office-Programmen beeinträchtigen.

Vor denselben Schwierigkeiten standen noch vor kurzem nicht nur Slavisten, sondern auch Vertreter anderer Philologien. Um diese und ähnliche Probleme zu beheben, wurde bereits Anfang der 1990er Jahre der Unicode-Standard eingeführt, der ab Version 5.1 um die kirchenslavischen Zeichen erweitert wurde[1]. Unicode ermöglicht es Sprachhistorikern und Mediävisten, kirchenslavische Texte in ein digitales Format zu überführen, das von allen Rechnern bzw. allen Betriebssystemen, derer sich Slavisten in der Regel bedienen, akzeptiert wird.

Wie oben angedeutet, wurde der als Digitalisierung bezeichnete Vorgang in der Slavistik vor der Erweiterung des Unicode-Standards um die historischen slavischen Schriftzeichen, aber auch bis in die jüngste Zeit vorwiegend mit selbstgebastelten idiosynkratischen Fonts durchgeführt. Dabei wurden fehlende Zeichen durch einzelne andere Zeichen oder Zeichenkombinationen von Unicode-Zeichen für nicht-kyrillische Schriften im Text ersetzt, um das äußere Bild oder den Sinn des gewünschten Buchstabens wiederzugeben.

[1] In Versionen 6.1 und 6.2 fanden sogar die am häufigsten gebrauchten supralinearen Buchstaben Berücksichtigung. Vgl. die Darstellung aller durch Unicode standardisierten kyrillischen Schriftzeichen unter http://de.wikipedia.//_Kyrillisch [4.10.13].

Ein bekanntes Beispiel für dieses Vorgehen ist die Erstellung eines deutschen Textes unter Verwendung der für das Englische angelegten Tastatur. Da auf der englischen Tastatur nicht alle Buchstaben des deutschen Alphabets vorhanden sind, wird mit Umschreibungen gearbeitet. So werden die Umlautbuchstaben *ä*, *ö* und *ü* durch Kombinationen mit *e* ersetzt: *ae*, *oe*, *ue*. Diese Umschreibungen sind den meisten Lesern vertraut und werden beim Lesen korrekt aufgelöst. In einer vergleichbaren Art gehen auch Slavisten vor, wenn sie fehlende Schriftzeichen wiedergeben wollen.

Damit die Texte, die mit Hilfe von Umschreibungen eingetippt wurden, auch für "nicht eingeweihte" Leser verständlich sind, erstellt man einen Font, üblicherweise einen Windows-Font, in dem die Code-Kombinationen aller Buchstaben und die entsprechenden Darstellungen dieser Buchstaben programmiert werden. Im genannten Beispiel sollen die Code-Kombinationen mit *e* – also *ae*, *oe*, *ue* – im Font von den Darstellungen *ä*, *ö*, *ü* abhängig gemacht werden: Wenn z. B. Microsoft Office den Code für *ue* liest, stellt das Programm als Ausgabe auf dem Bildschirm *ü* dar.

Dieses Vorgehen, das im Grunde genommen keine vertieften Kompetenzen im Bereich der Informatik erfordert, war und bleibt sehr verbreitet, so dass auf solche Weise seit Beginn der Computerära in der Editionsphilologie Dutzende Fonts bzw. Versionen bereits existierender Fonts erstellt wurden. Dabei einigte man sich für die Erstellung der benutzten Fonts nicht auf einen bestimmten Standard. Mehrere unterschiedliche Forschergruppen digitalisierten die zu edierenden Werke mit ihren eigenen, speziell für ihre Projekte zusammengestellten Schriftarten. Selbst innerhalb ein und derselben Forschergruppe hat man sich nicht immer über das Format des Fonts geeinigt. Die frei erfundenen Formate erschweren die Konvertierung der idiosynkratrisch digitalisierten Texte ins Unicode-Format wesentlich.

Ein weiteres Problem besteht darin, dass man bei der Digitalisierung der Werke bemüht war, das Layout der handschriftlichen Werke sowie graphische Besonderheiten der Schriftzeichen in der digitaler Version zu erhalten. Diese Informationen sind von großer Bedeutung für die Untersuchung unterschiedlicher mittelalterlicher Schreibschulen. Das heißt: Beim Konvertieren in Unicode darf die in den Editionen vorgenommene Formatierung für die Schriftart nicht verloren gehen.

In der vorgelegten Arbeit soll ein Tool präsentiert werden, das den Sprachhistorikern, Altphilologen und Mediävisten die Konvertierung der von ihnen vorbereiteten Editionen handschriftlicher Texte ins Unicode-Format erleichtern soll und sie beschleunigen könnte. Das Tool heißt *SlaVaComp-Konvertierer*. Es wurde im Rahmen des Projekts "SlaVaComp – COMPutergestützte Untersuchung von VAriabilität im KirchenSLAvischen" entwickelt, das in Zusammenarbeit zwischen dem Slavischen Seminar und dem Rechenzentrum

der Albert-Ludwigs-Universität Freiburg durchgeführt wird. Das unten dargestellte Programm ist nicht nur auf das Kirchenslavische beschränkt, sondern lässt sich auf alle Sprachen anwenden.

2. Tool-Beschreibung

Das Tool ist in der Programmiersprache C-Sharp auf .NET-Plattform geschrieben und ist ein Windows-Tool.

Als Input bekommt das Programm eine Word-Datei (**.docx*) und eine Excel-Datei (*defaultProfile.xlsx*). Beide Dateien sind im Office Open XML Format, das ab Office 2007 Standardformat ist. Um mit solchen Dateien arbeiten zu können, wird im Programm die Bibliothek Open XML SDK verwendet. Bei der Word-Datei handelt es sich um einen digitalisierten kirchenslavischen Text, der konvertiert werden soll. Bei der Excel-Datei handelt es sich um die Konvertierungstabelle, die die in der Word-Datei benutzten Fonts mit ihren Unicode-Pendants enthält. Ihr Format soll im Weiteren genauer betrachtet werden.

Das Vorgehen des Programms kann in groben Zügen folgendermaßen beschrieben werden: Das Konvertierungstool verwendet den XML-Code der Word-Datei. Dabei werden der Reihenfolge nach die Tags mit den Teiltexten und den dazugehörigen Informationen über den angewandten Font und die Formatierung extrahiert. Anhand dieser Informationen wird die entsprechende Konvertierungsfunktion aus der XML-Datei geholt, mit deren Hilfe der Text in Unicode konvertiert wird. Bei der Konvertierungsfunktion handelt es sich um einen Abbildungsprozess: Jede im Quell-Font verwendete Codierung der kirchenslavischen Buchstaben wird auf die entsprechende Unicode-Codierung abgebildet. Nach dem Konvertieren wird der extrahierte Tag zurückgespeichert. Dabei enthält er den konvertierten Text und Meta-Informationen über den neuen Unicode-Font und die Formatierung.

2.1. User-Interface

Das User-Interface des SlaVaComp-Konvertierungstools besteht aus einem Hauptmenü, einem Konvertierungsmenü und einer Konvertierungstabelle. Auf diese drei Bestandteile gehen wir im Folgenden ausführlicher ein.

2.1.1. Hauptmenü

Das Hauptmenü (s. Abb. 1) ist das erste Menü, das der Benutzer beim Starten des Programms zu sehen bekommt.

In diesem Menü wird die Konvertierung durchgeführt. Es enthält folgende Elemente:

(1) Pfad der Quelldatei (*Sourcefile*), die konvertiert werden soll. Wichtige Bedingung: Es muss eine Word-Datei im Office Open XML Format sein;

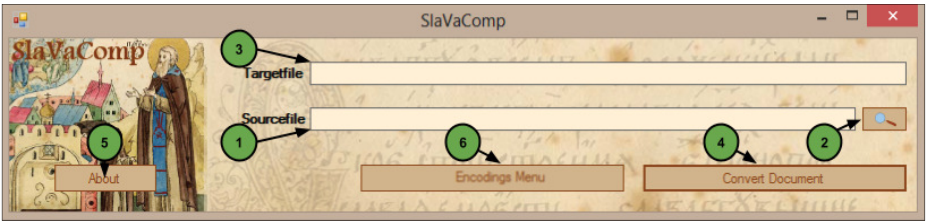

Abbildung 1. Hauptmenü

(2) Button zum Aufsuchen der Quelldatei auf dem Dateiträger (*Lupe als Symbol der Suchoption*);
(3) Pfad der resultierenden Unicode-Zieldatei (*Targetfile*): Nachdem die Quelldatei durch das Betätigen des Buttons (2) ausgesucht wurde, wird der Pfad der Zieldatei automatisch generiert. Die Zieldatei befindet sich in diesem Fall in demselben Ordner wie die Quelldatei und trägt denselben Namen, allerdings mit der Endung *_copy*;
(4) Button für die Konvertierungsoption (*Convert Document*): Durch das Betätigen dieses Buttons beginnt der Konvertierungsprozess;
(5) Button zum Öffnen der Tool-Beschreibung (*About*);
(6) Button zum Öffnen des Codierungsmenüs (*Encodings Menu*).

2.1.2. Codierungsmenü

Das Codierungsmenü (s. Abb. 2) ist das Menü der Codierungseinstellungen. Von hier aus kann man in der Konvertierungstabelle weitere Fonts hinzufügen bzw. die Tabelle bearbeiten.

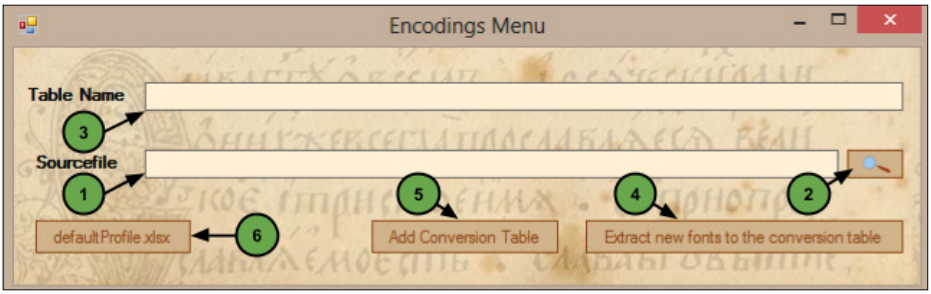

Abbildung 2. Codierungsmenü

Die Bestandteile des Codierungsmenüs sind:
(1) Pfad der Quelldatei (*Sourcefile*): Aus dieser Datei sollen fehlende Fonts (ohne die dazugehörigen Codierungen) in *defaultProfile.xlsx* hinzugefügt werden;

(2) Button zum Aufsuchen der Quelldatei auf dem Datenträger (*Lupe als Symbol der Suchoption*);
(3) Feld für den Namen der Tabelle (*Table Name*): In diese Tabelle werden alle gefundenen Fonts (ohne Codierungen) abgespeichert. Falls in diesem Feld nichts angegeben wird, bekommt die Tabelle den Namen *temp.xlsx* und wird in demselben Ordner wie die Quelldatei gespeichert;
(4) Button zum Extrahieren der neuen Fonts in die Konvertierungstabelle (*Extract new fonts to the conversion table*): Durch das Betätigen dieses Buttons wird die in (1) definierte Datei analysiert, wonach alle dort vorkommenden Fonts in die Tabelle eingetragen werden, die in (3) definiert ist. Außerdem werden in diese Tabelle alle gefundenen Fonts eingefügt, die in *defaultProfile.xlsx* noch fehlen. Dabei werden nur die entsprechenden Spalten für die Fonts erstellt. Die Spalten bleiben leer: Die dazugehörigen Codierungen müssen später manuell eingetragen werden.
(5) Button zum Integrieren der Tabelle in das Default Profile (*Add Conversion Table*): Mit diesem Button wird nach der bereits existierenden Konvertierungstabelle auf dem Datenträger gesucht, die dann in *defaultProfile.xlsx* integriert wird. Falls einer der Fonts in beiden Tabellen vorkommt, werden die Codierungen dieses Fonts zusammengeführt, so dass keine der Codierungen verloren geht.
(6) Button zum Öffnen des Default Profile (*defaultProfile.xlsx*): Mit diesem Button wird *defaultProfile.xlsx* geöffnet, das die Konvertierungstabelle enthält.

2.2. Konvertierungstabelle

Die Konvertierungstabelle (s. Abb. 3) ist eine Excel-Datei im Office Open XML Format. Nach der Installation des Programms befindet sie sich auf der Festplatte im Ordner *.../Programme(x86)/UniFreiburg/SlaVaComp/data* und trägt den Namen *defaultProfile.xlsx*. Es handelt sich dabei um eine Standard-Konvertierungstabelle für das Tool. In dieser Tabelle befinden sich mehrere Fonts mit den dazugehörigen Codierungen für die kirchenslavischen Buchstaben und ihre Unicode-Pendants. Diese Tabelle wurde im Rahmen des SlaVaComp-Projekts von mehreren Mitarbeitern des Slavischen Seminars der Universität Freiburg erstellt. Sie lässt sich erweitern. Mehr noch: Das Programm wurde so konzipiert, dass das Erweitern der Tabelle möglichst leicht fällt. Das Hauptziel ist dabei das Eintragen aller Non-Unicode-Fonts in diese Tabelle, so dass es dann möglich wäre, praktisch jedes digitalisierte kirchenslavische Werk ins Unicode-Format zu überführen. Die Datei hat dabei folgendes Format:

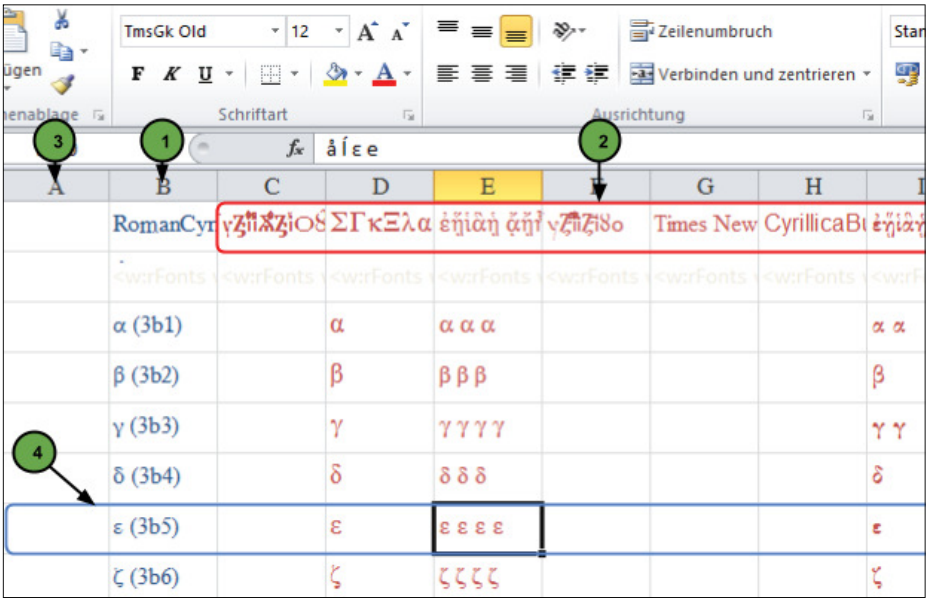

Abbildung 3. Konvertierungstabelle

(1) Die *Spalte B* enthält 2172 Unicode-Zeichen, darunter auch kirchenslavische, griechische, kyrillische und andere. Es handelt sich dabei um die Zielzeichen, d. h. diejenigen, in die konvertiert werden soll, also die Unicode-Zeichen. Bei dem eingangs erwähnten Beispiel würden dann in dieser Spalte *ä*, *ö*, *ü* mit ihrer Zahlendarstellung in Klammern stehen: *ä* – (0)0E4, *ö* – (0)0F6, *ü* – (0)0FC. In der ersten Zeile steht der Name des Unicode-Fonts, der im Programm verwendet wurde: *Roman Cyrillic*. In der zweiten Zeile stehen die automatisch konfigurierten Meta-Daten des Fonts: *<w:rFonts* \, die hier kaum zu sehen sind. Sie sind für den User nicht relevant.

(2) Alle Spalten ab der *Spalte C* sind für Non-Unicode-Fonts vorgesehen. Jede Spalte enthält in der ersten Zeile den Namen des Fonts, in der zweiten die automatisch erstellten Meta-Informationen. Alle weiteren Zeilen enthalten eine oder mehrere durch ein Leerzeichen verbundene Zeichenkombinationen, die mit den Codierungen der Unicode-Buchstaben im jeweiligen Non-Unicode-Font übereinstimmen.

(3) Die *Spalte A* ist für Kommentare wie etwa *Standardlateinisch, kombinierende diakritische Zeichen, Griechisch und Koptisch, Kyrillisch, Griechisch erweitert* usw. vorgesehen. Diese Kommentare, die den Ebenen in der Symbol-Tabelle entsprechen, erleichtern die Suche nach Schriftzeichen in der Konvertierungstabelle. Aber auch andere Kommentare können hier eingetragen werden.

(4) Jede Zeile steht für ein Unicode-Zeichen bzw. Fertigcode, d. h. Zeichen mit den zu ihm gehörenden Diakritika: <ἀ> (1F00) oder <ἓ> (1F14) und nicht <α> und <̓> oder <ε> und <̔>. Letzteres widerspräche vollkommen dem Unicode-Konzept: Einem Schriftzeichen, das eine lautliche Realisierung hat, soll nur ein Code entsprechen. Der Fertigcode ist in der zweiten Zelle (= *Spalte B*) zu sehen (vgl. Abb. 3). Alle weiteren Zellen sind für die Zeichen bzw. für die Zeichenkombinationen vorgesehen, die dieses Unicode-Zeichen in ihrem idiosynkratischen Font darstellen. In einigen Fonts wird ein Unicode-Zeichen durch mehrere Zeichenkombinationen dargestellt. In diesem Fall müssen die Zeichen, die eine Kombination bilden, ohne Leerzeichen miteinander verbunden werden, während die Kombinationen selbst durch ein Leerzeichen voneinander getrennt werden, damit das Programm mehrere Pendants für ein Zeichen unterscheiden kann.

(5) Falls sich ergeben sollte, dass die notwendigen Unicode-Zielzeichen in der Tabelle fehlen, besteht die Möglichkeit, am Dateiende, ab Zeile 2175, weitere Zeichen hinzuzufügen (s. Abb. 4). Dafür muss in die zweite Zelle (= *Spalte B*) einer neuen Zeile das benötigte Unicode-Zeichen und in alle weiteren Zellen die ihm entsprechenden Zeichenkombinationen eingetragen werden.

In einigen Fonts wird für die Darstellung mehrerer Buchstaben nur ein Zeichen benutzt. So gibt es zum Beispiel Fonts, in denen der kirchenslavische Digraph *oy* als Ligatur codiert wird. Obwohl dieser Digraph auch in Unicode als Fertigcode vorhanden ist (vgl. <Oʏ> 0478 und <oʏ> 0479), besteht das Unicode-Konsortium darauf, auf dessen Gebrauch zu verzichten und ihn durch die Kombination aus zwei Zeichen - <o> = 043E und <y> = 0443 – zu ersetzen. Falls im jeweiligen Font der Digraph als Ligatur codiert wurde, kann man diese Ligatur ebenfalls mit Hilfe der zusätzlichen Zeilen in die Kombination aus zwei Unicode-Zeichen umwandeln (s. Abb. 4).

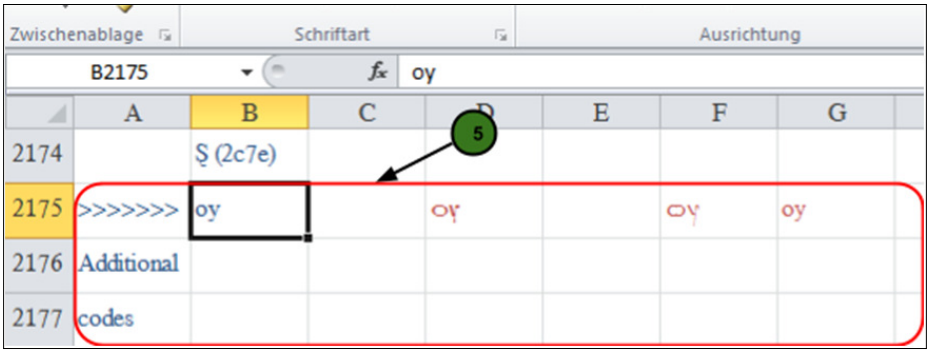

Abbildung 4. Konvertierungstabelle

2.3. Installation

Die einzige Installationskomponente ist eine Setup-Datei. Um das Programm zu installieren, muss man diese Datei ausführen und ihren Anweisungen folgen. Nach der Installation kann das Tool über eine Verknüpfung auf dem Desktop gestartet werden. Alternativ ist es möglich, das Programm direkt aus dem Ordner *.../Programme(x86)/UniFreiburg/SlaVaComp* zu starten. Systemvoraussetzungen für die Installation des Programms sind:

– Microsoft Office 2010 oder höher;
– Microsoft .NET Framework 4.5 oder höher;
– Open XML SDK 2.0 oder höher.

2.4. Fehlertoleranz

Beim Konvertieren können folgende Fehler auftreten:

(1) Das Programm erkennt den Font nicht, der in der Quelldatei benutzt wurde. Um festzustellen, ob dies der Fall ist, kann der Text nach dem Öffnen der resultierenden Datei mit dem Befehl *ctrl* + *f* nach der Zeichenkombination *fff* durchsucht werden. Falls man fündig wird, bedeutet dies, dass das Programm den Font des Textabschnittes zwischen *fff!* und *!* nicht erkennen kann. *fff* ist eine Fehlernachricht, die für einen nicht erkannten Font steht. Die Ausrufezeichen markieren die Grenzen des betroffenen Textabschnittes. Man kann das Problem lösen, indem man den Schritt 3 der Bedienung (s. u.) durchführt.

(2) Das Programm kann bestimmte Textabschnitte nicht konvertieren. Dieses Problem lässt sich feststellen, indem man die resultierende Datei nach der Fehlernachricht *ccc* durchsucht. Ausrufezeichen markieren die Grenzen des betroffenen Textabschnittes. Im Text kann *ccc* wie auch die Ausrufezeichen anders dargestellt werden (s. Abb. 5), und zwar im ursprünglichen idiosynkratischen Font des betroffenen Abschnitts. Dies wird dadurch verursacht, dass die Zeichen *c* oder *!* im zu konvertierenden Font anders codiert sind, was dann auch die Ausgabe verändert.

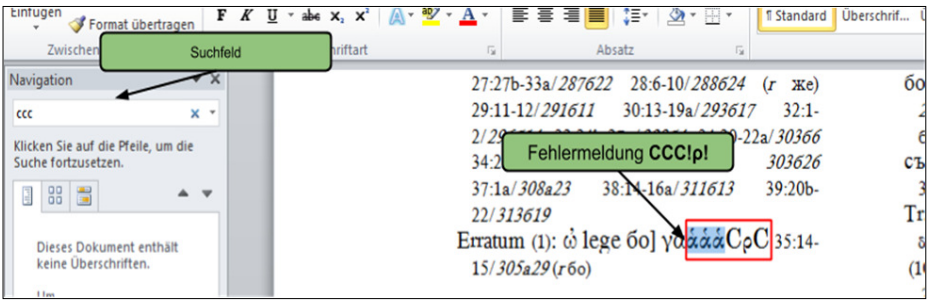

Abbildung 5. Fehlermeldung einer nicht erkannten Zeichenkombination

Der Grund für das Nicht-Konvertieren liegt meistens darin, dass in der Konvertierungstabelle *defaultProfile.xlsx* die entsprechenden Unicode-Codierungen, die im betroffenen Abschnitt zum Codieren der Buchstaben verwendet wurden, fehlen. Das Problem wird gelöst, indem die fehlenden Zeichencodierungen in die Konvertierungstabelle, in die Spalte mit dem betreffenden Font und in die Zeile des betreffenden Schriftzeichens, eingefügt werden.

2.5. Bedienung

Die allgemeine Vorgehensweise beim Bedienen des Programms lässt sich mit folgenden Schritten beschreiben:

Schritt 1: Starten des Programms. Suchen Sie mit 1 (s. Abb. 1) die zu konvertierende Datei auf und betätigen Sie Button 4. Dadurch wird die Konvertierungs-Routine gestartet. Warten Sie ab, bis das Programm den Prozess der Konvertierung beendet hat.

Schritt 2: Suche nach Font-Fehlern. Suchen Sie im Datenlaufwerk die resultierende Datei mit der Endung *_copy* auf und öffnen Sie sie mit Microsoft Office Word. Mit der Tastenkombination *ctrl + f* lässt sich das Suchmenü von Microsoft Office Word öffnen. Geben Sie *fff* in das Suchfeld ein. Falls es zu Treffern kommt, folgen Sie Schritt 3, sonst machen Sie weiter mit Schritt 4.

Schritt 3: Beseitigung von Font-Fehlern. Öffnen Sie das Codierungsmenü (s. Abb. 2). Suchen Sie mit Button 2 die Datei mit der Fehlermeldung *fff* auf und betätigen Sie Button 4. Nachdem das Extrahieren der fehlenden Fonts in die Konvertierungstabelle abgeschlossen ist, machen Sie weiter mit Schritt 4.

Schritt 4: Suche nach Codierungsfehlern. Suchen Sie auf dem Datenlaufwerk die resultierende Datei mit der Endung *_copy* auf und öffnen Sie sie mit Microsoft Office Word. Mit der Tastenkombination *ctrl + f* öffnet sich das Suchmenü von Microsoft Office Word. Geben Sie *ccc* in das Suchfeld ein. Falls es zu Treffern kommt, folgen Sie weiter Schritt 5. Falls kein Fehler gemeldet wird, ist die resultierende Datei vollständig ins Unicode-Format überführt.

Schritt 5: Beseitigung von Codierungsfehlern. Markieren Sie den ersten Buchstaben nach dem Ausrufezeichen im betroffenen Textabschnitt. Achten Sie dabei darauf, dass alle Zeichen markiert werden, mit denen dieser Buchstabe codiert wurde. Es empfiehlt sich die folgende Vorgehensweise:

- Fügen Sie unmittelbar vor dem zweiten Buchstaben ein Leerzeichen oder ein anderes gut erkennbares Zeichen ein;
- Bewegen Sie dann den Cursor vor das Ausrufezeichen;
- Bewegen Sie den Cursor mit der rechten Pfeiltaste einen Schritt nach rechts, so dass er vor dem ersten Buchstaben steht;
- Drücken Sie die *shift*-Taste und halten Sie sie;
- Bewegen Sie bei gedrückter *shift*-Taste den Cursor nach rechts, bis auch das Leerzeichen, das Sie vorher eingefügt haben, markiert wird;

– Gehen Sie anschließend einen Schritt zurück, so dass das Leerzeichen nicht mehr markiert ist;
– Kopieren Sie den markierten Abschnitt und fügen Sie ihn in die Konvertierungstabelle in die Spalte mit dem entsprechenden Font und in die Zeile mit dem entsprechenden Unicode-Zeichen ein. Gehen Sie dann zurück zu Schritt 4.

3. Fazit und Ausblick

Im vorliegenden Artikel wurde ein Tool namens SlaVaComp-Konvertierer vorgestellt. Es handelt sich dabei um ein Programm, das historisch arbeitenden Philologen und Linguisten helfen soll, digitalisierte Texte ohne Verlust der Formatierung in das Unicode-Format zu überführen. Eine wichtige Voraussetzung ist dabei, dass die Texte im Voraus im *docx*-Format gespeichert werden. Das Programm wurde für Werke in Kirchenslavisch und Alt- bzw. Mittelgriechisch konzipiert. Jedoch besteht die Möglichkeit, die Konvertierungstabelle mit weiteren Unicode-Zeichen zu vervollständigen und das Tool so für andere Non-Unicode-Fonts zu verwenden. Dies macht das Programm für alle historisch arbeitenden Philologen und nicht nur für Slavisten nutzbar. Das Ziel ist die Erstellung einer umfassenden Tabelle von Schriftzeichen, so dass in Zukunft der manuelle Aufwand beim Konvertieren minimiert wird. Der Autor bittet alle Benutzer des Tools um Benachrichtigung über Erweiterungen und eventuelle Probleme bei der Verwendung des SlaVaComp-Konvertierers (skilevic@googlemail.com).

Simon Skilevic
Albert Ludwig University of Freiburg

SlaVaComp Fonts Converter

Abstract

This paper presents a fonts converter that was developed as a part of the Freiburg project on historical corpus linguistics. The tool named SlaVaComp-Konvertierer converts Church Slavonic texts digitized with non-Unicode fonts into the Unicode format without any loss of information contained in the original file and without damage to the original formatting. It is suitable for the conversion of all idiosyncratic fonts—not only Church Slavonic—and therefore can be used not only in Palaeoslavistic, but also in all historical and philological studies.

Keywords

Unicode, font converter, editions of Church Slavonic texts, Church Slavonic alphabets and fonts, textual reuse, digitalizing, idiosyncratic fonts

Simon Skilevic, B. A.
Rechenzentrum der Albert-Ludwigs-Universität Freiburg
Hermann-Herder-Str.10
D-79104 Freiburg im Breisgau
Deutschland / Germany
skilevic@googlemail.com